Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 1996		**00-00-1996 to 00-00-1996**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Approaches in MET (Multi-lingual Entity Task)	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **BBN Systems and Technologies,70 Fawcett Street,Cambridge,MA,02138**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996. Sponsored by the Defense Advanced Research Projects Agency.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**2**	

APPROACHES IN MET (MULTI-LINGUAL ENTITY TASK)

Damaris Ayuso, Daniel Bikel, Tasha Hall, Erik Peterson, Ralph Weischedel
BBN Systems and Technologies
70 Fawcett Street, Cambridge, MA 02138
weischedel@bbn.com
617-873-3496

Patrick Jost
Financial Crimes Enforcement Network
(FinCEN)
2070 Chain Bridge Road
Vienna, VA 22182
703-905-3648

1. TWO APPROACHES

BBN and FinCEN participated jointly in the Spanish language task for MET. BBN also participated in Chinese. We also fielded two approaches. The first approach is pattern based and has an architecture as shown in Figure 1. This approach was applied to both Chinese and Spanish. The algorithms (rectangles in the Figure) were used in the two languages; the only component difference was the New Mexico State University segmenter, used to find the word boundaries in Chinese. The components common to both languages are the message reader, which dealt with the input format and SGML conventions via a declarative format description; the part-of-speech tagger (BBN POST); a lexical pattern matcher driven by knowledge bases of patterns and lexicons specific to each language; and the SGML annotation generator. While not shown in Figure 1, an alias prediction algorithm was shared by both languages, using patterns unique to each language.

A second approach based on statistical learning was used to create a learned Spanish namefinder. One component is a training module that learns to recognize the MET categories from examples. The understanding module uses the model developed from training to predict the MET categories in new input sentences. Data annotated with the correct answers was provided by the government in its training materials. In addition, we annotated some additional data. The current probability model is a hidden Markov model (HMM) which is more complex than is typically used in part-of-speech tagging and is therefore more general.

2. CHALLENGES AND STRENGTHS IN OUR APPROACH TO CHINESE

One of the challenges in processing Chinese is the difficulty of word segmentation. Segmentation in Chinese seems more difficult than in Japanese. With Japanese, changes in the character sets used in running text can be used to detect many of the word boundaries.

The use of the part-of-speech tagger was both a strength and a weakness in Chinese. The part-of-speech labels proved useful in finding boundaries such as those between organization names and text which is not one of the MET categories. However, part-of-speech labeling in Chinese is more of a challenge than in the other languages because of two factors:

- Chinese has very little inflection and no capitalization, thereby offering less evidence to predict the category of an unknown word.

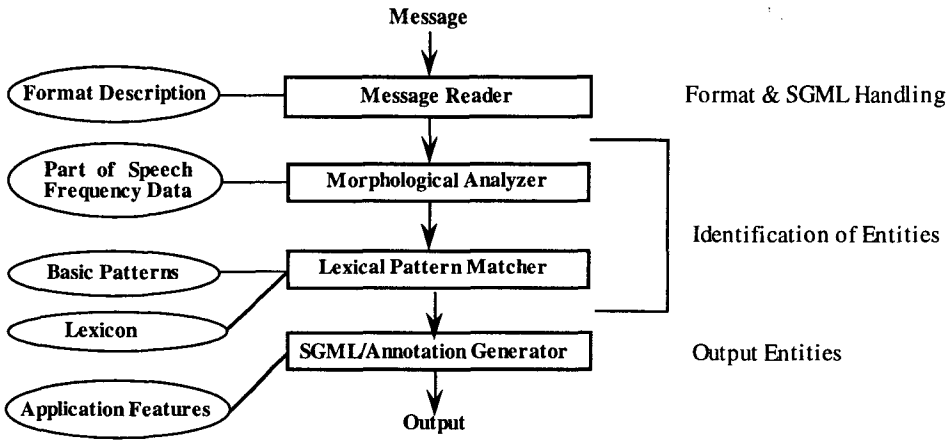

Figure 1: IdentiFinder System Architecture: *Rectangles represent domain-independent, language-independent algorithms; ovals represent knowledge bases*

- Given that there was not a large dictionary of Chinese words with parts-of-speech, a high percentage of words in the text were unknown.

Another strength and challenge in Chinese is the fact that several of the categories are interrelated. For instance, locations often mark the start of an organization name and persons may start an organization name. In addition, different categories will occur contiguously, so that correctly recognizing a category is needed to locate the others. For example,. a location name, a title of a person, and a person name often will co-occur. This creates a challenge in getting started since several of the patterns look for distributed categories. The strength is that once significant progress is made in one, such as location names, it can contribute to improved performance in the other categories.

The final general challenge is represented by the lack of available linguistics resources for Chinese.

3. CHALLENGES AND STRENGTHS IN SPANISH

3.1 Using manually constructed patterns

One of the challenges was self-imposed: because we were interested in seeing how far the technology could go without purchased linguistics resources, we restricted ourselves to using only prelinguistics resources. Some of the techniques we used are therefore applicable in all languages where significant amounts of online text are available. Patrick Jost was very effective in mining available online data to find very large lists of person names, critical vocabulary items, and organization names. A second challenge was that we had very little effort to devote to the manual system in Spanish; in fact, after a certain point there was insufficient effort available to track the evolving set of guidelines for Spanish. One strength in the effort was that the presence of lower case words in Spanish names (and the generally unreliable use of capitalization in the names) was straightforwardly handled by the patterns and did not pose a difficulty as we would have anticipated.

3.2 Using a Learned System

There are several pleasant surprises corresponding to strengths in the learned system as applied to Spanish. First the learned system could be retrained in a matter of five or ten minutes. Therefore, changes to the model could be quickly tested. The fact that the government released the revised training data very late in the cycle of MET did not pose a problem, since the system could be retrained so quickly with the updated training data.

The learned system and model we used proved to be highly portable to a new language. The original training and understanding modules were not completed until the first half of March. Results were very positive in English. When we first trained and tested the same model in Spanish, the results were so encouraging that we decided in April to enter the learned system in MET.

The third strength we found was the use of contextual probabilities to predict from the previous word and previous category the likelihood of the next word and the next category.

The major challenge is to make the resulting large statistical model more understandable by humans, so that intuitions can be used to improve it.

4. LESSONS LEARNED

We learned the following lessons:

- High performances are possible using one approach across several languages.
- Text can be mined using simple techniques (such as regular expression patterns) to effectively find critical vocabulary items.
- The gap between manually constructed systems using patterns and learned systems is shrinking dramatically.
- Probabilistic, learned approaches can be developed in a short amount of time.
- Probabilistic finite state models, which had been previously successful in continuous speech recognition and in part-of-speech tagging, can be applied successfully to multilingual entity finding.

5. ACKNOWLEDGMENTS

The work reported here was supported in part by the Defense Advanced Research Projects Agency; technical agents for part of the work were Rome Laboratory under contract number F30602-95-C-0111 and Fort Huachucha under contract number DABT63-94-C-0062. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies either expressed or implied, of the Defense Advanced Research Projects Agency or the United States Government. Sarah Law and Rusty Bobrow contributed to the work on Spanish, Scott Miller and Richard Schwartz contributed to the ideas of the learned approach.